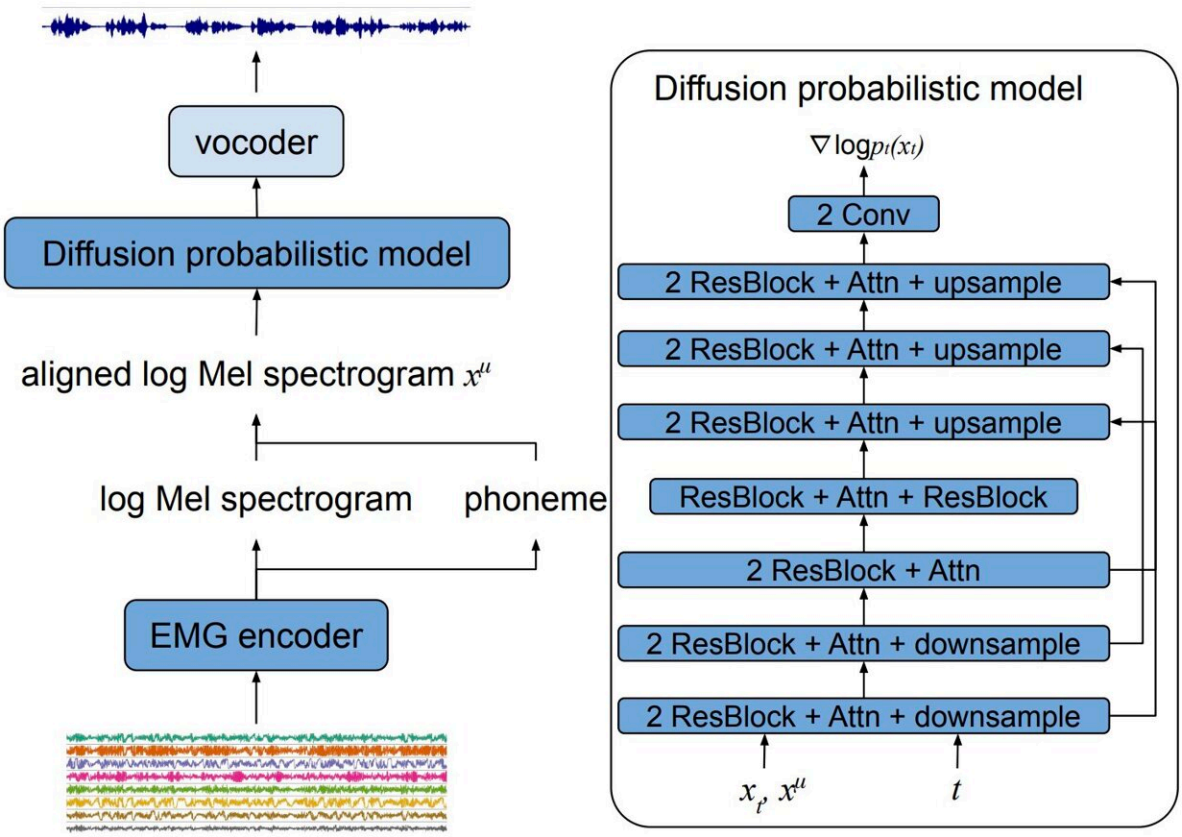
# A new model to produce more natural synthesized speech

May 27 2024, by Ingrid Fadelli




The proposed Diff-ETS framework for ETS. The deep blue blocks are trainable and the light blue block of the vocoder is frozen. ResBlock: Residual blocks, Attn: Attention, Conv: Convolutional layers. Credit: Ren et al

Recent technological advances are enabling the development of computational tools that could significantly improve the quality of life of individuals with disabilities or sensory impairments. These include so-called electromyography-to-speech (ETS) conversion models, designed to convert electrical signals produced by skeletal muscles into speech.

Researchers at University of Bremen and SUPSI recently introduced Diff-ETS, a model for ETS conversion that could produce more natural synthesized speech. This model, introduced in a paper posted to the preprint server *arXiv*, could be used to develop new systems that allow people who are unable to speak, such as patients who underwent a laryngectomy (a surgery to remove part of the human voice box), to communicate with others.

Most previously introduced techniques for ETS conversion have two key components: an EMG encoder and a vocoder. The electromyography (EMG) encoder can convert EMG signals into acoustic speech features, while the vocoder uses these speech features to synthesize speech signals.

"Due to an inadequate amount of available data and noisy signals, the synthesized speech often exhibits a low level of naturalness," Zhao Ren, Kevin Scheck and their colleagues wrote in their paper. "In this work, we propose Diff-ETS, an ETS model which uses a score-based diffusion probabilistic model to enhance the naturalness of synthesized speech. The diffusion model is applied to improve the quality of the acoustic features predicted by an EMG encoder."

In contrast with many other ETS conversion models developed in the past, consisting of an encoder and vocoder, the researchers' model has three components, namely an EMG encoder, a diffusion probabilistic model and a vocoder. The diffusion probabilistic model, the second of these components, is thus a new addition, which could result in more

natural synthesized speech.

Ren, Scheck and their colleagues trained the EMG encoder to predict a so-called log Mel spectrogram (i.e., a visual representation of audio signals) and phoneme targets from EMG signals. The diffusion probabilistic model, on the other hand, was trained to enhance log Mel spectrograms, while the pre-trained vocoder can translate this spectrogram into synthesized speech.

The researchers evaluated the Diff-ETS model in a series of tests, comparing it with a baseline ETS technique. Their findings were highly promising, as the speech it synthesized was more natural and human-like than that produced by the baseline method.

"In our experiments, we evaluated fine-tuning the diffusion model on predictions of a pre-trained EMG encoder, and training both models in an end-to-end fashion," Ren, Scheck and their colleagues wrote in their paper. "We compared Diff-ETS with a baseline ETS model without diffusion using objective metrics and a listening test. The results indicated the proposed Diff-ETS significantly improved speech naturalness over the baseline."

In the future, the ETS conversion model developed by this team of researchers could be used to develop better technologies for the artificial generation of audible speech. These systems could allow people who are unable to speak to express their thoughts out loud, facilitating their interaction with others.

"In future efforts, one can reduce the number of model parameters using various methods, e. g., model compression and knowledge distillation, thereby generating speech samples in real-time," the researchers wrote. "Moreover, a diffusion model can be trained together with the encoder and vocoder for further enhancing the speech quality."

**More information:** Zhao Ren et al, Diff-ETS: Learning a Diffusion Probabilistic Model for Electromyography-to-Speech Conversion, *arXiv* (2024). DOI: 10.48550/arxiv.2405.08021